

Smoothed analysis of componentwise condition numbers for sparse matrices

Dennis Cheung
United International College
Tang Jia Wan
Zhuhai, Guandong Province
P.R. of CHINA
e-mail: dennisc@uic.edu.hk

Felipe Cucker ^{*}
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon
HONG KONG
e-mail: macucker@cityu.edu.hk

Abstract. We perform a smoothed analysis of the componentwise condition numbers for determinant computation, matrix inversion, and linear equations solving for sparse $n \times n$ matrices. The bounds we obtain for the expectations of the logarithm of these condition numbers are, in all three cases, of the order $\mathcal{O}(\log n)$. As a consequence, small bounds on the smoothed loss of accuracy for triangular linear systems follow.

1 Introduction

The most commonly used solver of linear systems of equations, Gaussian elimination, reduces the input system $Ax = c$ to a system $Lx = b$ with L lower triangular (and same solution x). Then, it solves the latter by forward substitution. As a consequence, triangular systems of equations are routinely solved by computer.

Almost on every occasion, the accuracy of the computed solution is very high. Yet, the reasons for this accuracy have been dodging researchers for quite a while. In the early 1960s J.H. Wilkinson noted that “In practice one almost invariably finds that if L is ill-conditioned, so that $\|L\|\|L^{-1}\| \gg 1$, then the computed solution of $Lx = b$ (or the computed inverse) is far more accurate than [what forward stability analysis] would suggest” [11, p. 105]. To make things worse, ill-conditioned matrices L in the sense above, appeared to be ubiquitous. This was explained by by D. Viswanath and

^{*}Partially supported by GRF grant CityU 100808

N. Trefethen in [9]. Indeed, if L_n denotes a random triangular $n \times n$ matrix (whose entries are independent standard Gaussian random variables) and $\kappa_n = \|L_n\| \|L_n^{-1}\|$ is its condition number (which is a positive random variable) then, the main result in [9] shows that

$$\sqrt[n]{\kappa_n} \rightarrow 2 \quad \text{almost surely}$$

as $n \rightarrow \infty$. A straightforward consequence of this result is that the expected value of $\log \kappa_n$ satisfies $\mathbb{E}(\log \kappa_n) = \Omega(n)$.

Putting all the above together we can describe the situation as follows:

Triangular systems of equations are generally solved to high accuracy in spite of being, in general, ill-conditioned.

In 1989 N. Higham [3] pointed out that the backward error analysis given by Wilkinson for forward substitution yields (small) *componentwise* bounds on the perturbed matrix. One can therefore deduce small forward error bounds for these solutions if the *componentwise condition number* $c(L, b)$ of the pair (L, b) —instead of $\kappa(L)$ — is small. In a recent paper [2] we showed that this is the case for random triangular matrices L . Here ‘random’ means that the entries of L are i.i.d. standard random variables. This result provides an explanation of the high accuracy achieved in general by forward substitution.

In the last decade, however, the suitability of this *average analysis* to reflect performance of algorithmic practice was questioned. The objection raised is that the probability distribution underlying these analyses —usually, a centered isotropic Gaussian— is chosen because of technical reasons and not because it models “the real world.” Because of this, it may well happen that the resulting estimates are too optimistic, just as worst-case analysis is often claimed to be too pessimistic. The proposed alternative, *smoothed analysis*, interpolates between worst-case and average analyses and typically studies, for a function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, the quantity

$$\sup_{\bar{a} \in \mathbb{R}^p} \mathbb{E}_{a \sim N(\bar{a}, \sigma^2 \text{Id})} f(a).$$

Here $N(\bar{a}, \sigma^2 \text{Id})$ denotes the normal distribution centered at \bar{a} and with covariance matrix $\sigma^2 \text{Id}$, where Id is the identity matrix. In case f is homogeneous (i.e., $f(\lambda a) = f(a)$ for all $\lambda \neq 0$) it is common to scale the covariance matrix and study

$$\sup_{\bar{a} \in \mathbb{R}^p} \mathbb{E}_{a \sim N(\bar{a}, \sigma^2 \|\bar{a}\| \text{Id})} f(a).$$

or, equivalently,

$$\sup_{\|\bar{a}\|=1} \mathbb{E}_{a \sim N(\bar{a}, \sigma^2 \text{Id})} f(a).$$

In this case, the interpolation mentioned above comes from the fact that when $\sigma = 0$ the expression above reduces to the worst-case of f and when $\sigma \rightarrow \infty$ one approaches the usual average analysis. We won't elaborate here on the virtues of smoothed analysis. The interested reader can find expositions of these virtues in [4, 5, 6, 7] or [1, §2.2.7]. We will instead proceed to state the main results of this paper. For a matrix A we define the max norm

$$\|A\|_{\max} = \max_{ij} |a_{ij}|.$$

Theorem 1. *Let \mathcal{T} denote the set of $n \times n$ lower triangular matrices. Let $\bar{L} \in \mathcal{T}$ and $\bar{b} \in \mathbb{R}^n$ be such that $\|\bar{L}\|_{\max} \leq 1$ and $\|\bar{b}\|_{\infty} \leq 1$. For $L \in \mathcal{T}$ and $b \in \mathbb{R}^n$ let $\mathbf{c}(L, b)$ denote the componentwise condition number, for the problem of linear equation solving, of the pair (L, b) . Then, for any real number $t > n(n+1)$ we have*

$$\text{Prob}_{(L,b) \sim N_{\mathcal{T}}((\bar{L}, \bar{b}), \sigma^2 \text{Id})} \{\mathbf{c}(L, b) > t\} \leq \left(\frac{1+\sigma}{\sigma} \right) \left(\frac{n^3(n+1)^2}{t - n(n+1)} \right) \sqrt{\frac{2}{\pi}}$$

and, for any $\beta > 1$,

$$\mathbb{E}_{(L,b) \sim N_{\mathcal{T}}((\bar{L}, \bar{b}), \sigma^2 \text{Id})} (\log_{\beta}(\mathbf{c}(L, b))) \leq \log_{\beta} \left(\frac{1+\sigma}{\sigma} \right) + 5 \log_{\beta}(n) + \frac{2.65}{\ln \beta}.$$

The subindex \mathcal{T} in $N_{\mathcal{T}}((\bar{L}, \bar{b}), \sigma^2 \text{Id})$ is meant to denote that L is triangular. That is, the only entries of L which are drawn from the Gaussian $N((\bar{L}, \bar{b}), \sigma^2 \text{Id})$ are those in its lower part.

This theorem has immediate consequences for the accuracy of forward substitution. Recall (or look at the *Overture* chapter in [1] for a primer if you are not familiar with round-off analysis), a finite precision algorithm with *machine precision* $\varepsilon_{\text{mach}}$ rounds-off all the real numbers z occurring in the execution to a rational (floating point) number \tilde{z} satisfying

$$\text{RelError}(z) := \frac{|\tilde{z} - z|}{|z|} \leq \varepsilon_{\text{mach}}$$

(we agree this equality to hold if $z = \tilde{z} = 0$). This means that the approximation \tilde{z} has $\log_{10}(\frac{1}{\varepsilon_{\text{mach}}})$ correct (significant) digits¹.

¹All our discussion holds as well for bits, instead of digits. The modifications required are trivial.

If we solve a system $Lx = b$ with a finite precision machine we obtain an approximation \tilde{x} of the solution x . A (componentwise) extension of the notion above measures the relative error of this approximation by

$$\text{RelError}(x) := \max_{i \leq n} \text{RelError}(x_i).$$

Again, $\log_{10}(\text{RelError}^{-1}(x))$ provides a lower bound on the number of correct digits for *all* the components of x and hence the *loss of precision* in the computation of x —i.e., the initial precision of our data measured in number of correct digits minus the precision of the computed outcome measured in the same manner—is

$$\text{LoP}(x) := \log_{10}(\varepsilon_{\text{mach}}^{-1}) - \log_{10}(\text{RelError}^{-1}(x)).$$

Note that if L is singular then x is not well-defined or may not exist. In this case we take, by convention, $\text{LoP}(x) = \infty$. The following result provides a smoothed analysis of this quantity for forward substitution with finite precision.

Corollary 1. *Assume we solve systems $Lx = b$ using forward substitution. Then, for all $\bar{L} \in \mathcal{T}$ and $\bar{b} \in \mathbb{R}^n$ with $\|\bar{L}\|_{\max} \leq 1$ and $\|\bar{b}\|_{\infty} \leq 1$ we have*

$$\mathbb{E}(\text{LoP}(x)) = \log_{10}\left(\frac{1+\sigma}{\sigma}\right) + 5 \log_{10} n + \log_{10}(\log_2 n) + 1.452 + o(1).$$

Here $(L, b) \sim N_{\mathcal{T}}((\bar{L}, \bar{b}), \sigma^2 \text{Id})$ and $o(1)$ is a quantity that tends to zero with $\varepsilon_{\text{mach}}$.

2 Preliminaries

2.1 Componentwise condition numbers

Condition numbers measure the worst-case magnification in the computed outcome of a small perturbation in the data. As originally introduced by Turing [8], or von Neumann and Goldstine [10], they were *normwise* in the sense that data perturbation and outcome’s error were measured using norms (in the space of data and outcomes respectively). In contrast, *componentwise* condition numbers measure both of them componentwise.

For both data perturbation and output error, the error is measured in a relative manner. Because of this, the following form of “distance” function

(it is not a distance as is not symmetric) will be useful to define componentwise condition numbers. For points $u, v \in \mathbb{R}^p$ we define $\frac{u}{v} = (w_1, \dots, w_p)$ with

$$w_i = \begin{cases} u_i/v_i & \text{if } v_i \neq 0 \\ 0 & \text{if } u_i = v_i = 0 \\ \infty & \text{otherwise.} \end{cases}$$

Then we define

$$d(u, v) := \left\| \frac{u - v}{v} \right\|_{\infty}.$$

Note that, if $d(u, v) < \infty$,

$$d(u, v) := \min\{\nu \geq 0 \mid |u_i - v_i| \leq \nu|v_i| \text{ for } i = 1, \dots, p\}.$$

For $\delta > 0$ and $a \in \mathbb{R}^p$ we denote $\mathcal{S}(a, \delta) = \{x \in \mathbb{R}^p \mid d(x, a) \leq \delta\}$.

Let $\mathcal{D} \subseteq \mathbb{R}^p$, $F : \mathcal{D} \rightarrow \mathbb{R}^q$ be a continuous mapping, and $a \in \mathcal{D}$ be such that $a_j \neq 0$ for $j = 1, \dots, q$. Then the *componentwise condition number* of F at a is

$$\mathbf{c}^F(a) := \lim_{\delta \rightarrow 0} \sup_{\substack{x \in \mathcal{S}(a, \delta) \\ x \neq a}} \frac{d(F(x), F(a))}{d(x, a)}. \quad (1)$$

It is not difficult to see that

$$\mathbf{c}^F(a) = \max_{j \leq q} \mathbf{c}^{F_j}(a) \quad (2)$$

where $\mathbf{c}^{F_j}(a)$ denotes the componentwise condition number of a for the j th component F_j of F . We will systematically use this form in the rest of this paper.

2.2 Sparse matrices

In all what follows, for $n \in \mathbb{N}$, we denote the set $\{1, \dots, n\}$ by $[n]$.

We denote by \mathcal{M} the set of $n \times n$ real matrices and by Σ its subset of singular matrices. Also, for a subset $S \subseteq [n]^2$ we denote

$$\mathcal{M}_S = \{A \in \mathcal{M} \mid \text{if } (i, j) \notin S \text{ then } a_{ij} = 0\}.$$

Matrices in \mathcal{M}_S for some $S \neq [n]^2$ (i.e. matrices with a fixed pattern of zeros) are said to be *sparse*. The set S is said to be *admissible* if \mathcal{M}_S contains some invertible matrix.

In the rest of this paper, for non-singular matrices A, A' , we denote their inverses by Γ, Γ' , respectively. Also, we denote by A_{ij} the sub-matrix of A obtained by removing from A its i th row and its j th column.

The technical results below, Theorems 2, 3 and 4, are proved in the general context of sparse matrices. Besides triangular matrices, these results apply to other classes of sparse matrices such as, for instance, tridiagonal matrices.

2.3 Smoothed analysis

Let $\sigma > 0$ be a fixed number, $S \subset [n]^2$ be admissible and $\bar{A} = (\bar{a}_{ij}) \in \mathcal{M}_S$. Extending the notation we used in the Introduction, we will write $A \sim N_S(\bar{A}, \sigma^2 \|\bar{A}\|_{\max} \text{Id})$ to denote that the entry a_{ij} of A , with $(i, j) \in S$, is a random variable with distribution $N(\bar{a}_{ij}, \sigma^2 \|\bar{A}\|_{\max})$, whereas the entries a_{ij} with $(i, j) \notin S$ are zero.

In this paper we will only be concerned, for a random sparse matrix A as above, with the componentwise condition number of A with respect to a few problems. All these condition numbers being, as functions, homogeneous of degree 0, we will systematically consider, without loss of generality, the center \bar{A} of the distribution to satisfy $\|\bar{A}\|_{\max} = 1$ (or, more generally and for convenience, $\|\bar{A}\|_{\max} \leq 1$) and therefore, we will take $\sigma^2 \text{Id}$ as covariance matrix in our distributions.

3 Preliminary results

We prove in this section some bounds on one-dimensional Gaussian random variables as well as a proposition on the expectation of positive random variables. The main results of the paper will easily follow from them.

Proposition 1. *Let $\mu, \varsigma > 0$ and $t > 1$ be fixed numbers. Let $X \sim N(\mu, \varsigma^2)$ be a normal distributed random variable. Then*

$$\text{Prob}\{|X| > t|X + 1|\} < \left(\frac{|\mu| + \varsigma}{\varsigma} \right) \left(\frac{1}{t-1} \right) \sqrt{\frac{2}{\pi}}.$$

The proof of Proposition 1 proceeds through a sequence of lemmas.

Lemma 1. *Let $\mu \in \mathbb{R}$ and $\varsigma > 0$ be fixed numbers. Let $X \sim N(\mu, \varsigma^2)$ be a normal distributed random variable. Then*

$$\text{Prob}\{1 < X < 1 + \varepsilon\} \leq \frac{\varepsilon}{\varsigma} \sqrt{\frac{1}{2\pi}}.$$

Proof. Since $X \sim N(\mu, \varsigma^2)$

$$\begin{aligned} \text{Prob}\{1 < X < 1 + \varepsilon\} &= \frac{1}{\varsigma} \sqrt{\frac{1}{2\pi}} \int_1^{1+\varepsilon} e^{-\frac{(x-\mu)^2}{2\varsigma^2}} dx \\ &\leq \frac{1}{\varsigma} \sqrt{\frac{1}{2\pi}} \int_1^{1+\varepsilon} 1 dx = \frac{\varepsilon}{\varsigma} \sqrt{\frac{1}{2\pi}}. \end{aligned}$$

□

Lemma 2. Let $\mu \in \mathbb{R}$, $\varsigma > 0$. Let $X \sim N(\mu, \varsigma^2)$ be a Gaussian random variable. Then

$$\text{Prob}\{1 < X < 1 + \varepsilon\} \leq \varepsilon \left(\frac{|\mu| + \varsigma}{\varsigma} \right) \sqrt{\frac{1}{2\pi}}.$$

Proof. We first assume that $\mu \geq 0$. Let $r = \frac{\varsigma}{\mu}$ and $f : \mathbb{R} \rightarrow \mathbb{R}$ be given by

$$f(m) = \frac{1}{mr} \sqrt{\frac{1}{2\pi}} \int_1^{1+\varepsilon} e^{-\frac{(x-m)^2}{2m^2r^2}} dx$$

so that $\text{Prob}\{1 < X < 1 + \varepsilon\} = f(\mu)$. By doing the change of variables $u = \frac{x-m}{mr\sqrt{2}}$ we obtain,

$$\begin{aligned} f(m) &= \sqrt{\frac{1}{\pi}} \int_{\frac{1-m}{mr\sqrt{2}}}^{\frac{1+\varepsilon-m}{mr\sqrt{2}}} e^{-u^2} du \\ &= \sqrt{\frac{1}{\pi}} \left(\int_0^{\frac{1+\varepsilon-m}{mr\sqrt{2}}} e^{-u^2} du - \int_0^{\frac{1-m}{mr\sqrt{2}}} e^{-u^2} du \right) \end{aligned}$$

and, hence,

$$f'(m) = \sqrt{\frac{1}{\pi}} \left(\frac{d}{dm} \int_0^{\frac{1+\varepsilon-m}{mr\sqrt{2}}} e^{-u^2} du - \frac{d}{dm} \int_0^{\frac{1-m}{mr\sqrt{2}}} e^{-u^2} du \right).$$

Let $v = \frac{1+\varepsilon-m}{mr\sqrt{2}}$ and $w = \frac{1-m}{mr\sqrt{2}}$. Then

$$f'(m) = \sqrt{\frac{1}{\pi}} \left(\frac{d}{dm} \int_0^v e^{-u^2} du - \frac{d}{dm} \int_0^w e^{-u^2} du \right).$$

By the chain rule and the Fundamental Theorem of Calculus,

$$\begin{aligned} f'(m) &= \sqrt{\frac{1}{\pi}} \left(\frac{dv}{dm} \cdot \frac{d}{dv} \int_0^v e^{-u^2} du - \frac{dw}{dm} \cdot \frac{d}{dw} \int_0^w e^{-u^2} du \right) \\ &= \sqrt{\frac{1}{\pi}} \left(\frac{dv}{dm} e^{-v^2} - \frac{dw}{dm} e^{-w^2} \right). \end{aligned} \tag{3}$$

We now use that

$$\frac{dv}{dm} = \frac{-(1+\varepsilon)}{m^2 r \sqrt{2}} \quad \text{and} \quad \frac{dw}{dm} = \frac{-1}{m^2 r \sqrt{2}}$$

to deduce from (3) that

$$\begin{aligned} -m^2 r \sqrt{2\pi} f'(m) &= e^{-v^2} - (1+\varepsilon)e^{-w^2} \\ &= e^{-\frac{(1+\varepsilon-m)^2}{2m^2 r^2}} - (1+\varepsilon)e^{-\frac{(1-m)^2}{2m^2 r^2}}. \end{aligned} \quad (4)$$

Let m_* be such that

$$f(m_*) = \sup_{m \geq 0} f(m).$$

Since $\lim_{m \rightarrow \infty} f(m) = \lim_{m \rightarrow 0} f(m) = 0$ we deduce that $f'(m_*) = 0$. Equation (4) evaluated at m_* then yields

$$e^{-\frac{(1-m_*)^2}{2m_*^2 r^2}} = (1+\varepsilon)e^{-\frac{(1+\varepsilon-m_*)^2}{2m_*^2 r^2}}$$

which elementary computations show equivalent to

$$\varepsilon^2 + 2\varepsilon(1-m_*) = 2m_*^2 r^2 \ln(1+\varepsilon).$$

Since $\ln(x) \leq x - 1$ for all $x > 0$ this last equality implies that

$$\varepsilon + 2 - 2m_* \leq 2m_*^2 r^2$$

which in turn implies, since $\varepsilon > 0$,

$$r^2 m_*^2 + m_* - 1 > 0.$$

Solving this quadratic inequality we deduce that either

$$2r^2 m_* > -1 + \sqrt{1 + 4r^2}$$

or

$$2r^2 m_* < -1 - \sqrt{1 + 4r^2}$$

but we can reject the latter since $m_* \geq 0$. The former inequality can also be written as

$$m_* r > \frac{-1 + \sqrt{1 + 4r^2}}{2r} = \frac{4r^2}{2r(1 + \sqrt{1 + 4r^2})} = \frac{2r}{1 + \sqrt{1 + 4r^2}} \geq \frac{r}{1 + r}.$$

Let $Y \sim N(m_*, m_* r)$. Using Lemma 1 and this inequality we deduce that

$$f(m_*) = \text{Prob}\{1 < Y < 1+\varepsilon\} \leq \frac{\varepsilon}{m_* r} \sqrt{\frac{1}{2\pi}} \leq \varepsilon \frac{1+r}{r} \sqrt{\frac{1}{2\pi}} = \varepsilon \frac{\mu + \varsigma}{\varsigma} \sqrt{\frac{1}{2\pi}}.$$

The statement (for the case $\mu \geq 0$) now follows since

$$\text{Prob}\{1 < X < 1 + \varepsilon\} = f(\mu) \leq f(m_*).$$

We next deal with the case $\mu < 0$. Since $X \sim N(\mu, \varsigma^2)$,

$$\text{Prob}\{1 < X < 1 + \varepsilon\} = \frac{1}{\varsigma} \sqrt{\frac{1}{2\pi}} \int_1^{1+\varepsilon} e^{-\frac{(x-\mu)^2}{2\varsigma^2}} dx. \quad (5)$$

Let $Y \sim N(-\mu, \varsigma^2)$. Then

$$\text{Prob}\{1 < Y < 1 + \varepsilon\} = \frac{1}{\varsigma} \sqrt{\frac{1}{2\pi}} \int_1^{1+\varepsilon} e^{-\frac{(x+\mu)^2}{2\varsigma^2}} dx. \quad (6)$$

Since $(x + \mu)^2 < (x - \mu)^2$ for all $x \in (1, 1 + \varepsilon)$, using (5) and (6) we obtain

$$\text{Prob}\{1 < Y < 1 + \varepsilon\} \geq \text{Prob}\{1 < X < 1 + \varepsilon\}.$$

The result now follows since, by the first case above, the claimed bound holds for Y . \square

Proof of Proposition 1. We have

$$\begin{aligned} |X| > t|X + 1| &\iff X^2 > t^2(X + 1)^2 \\ &\iff (t^2 - 1)X^2 + 2t^2X + t^2 < 0 \\ &\iff \frac{-t}{t-1} < X < \frac{-t}{t+1} \\ &\iff \frac{t+1}{t-1} > \left(-\frac{t+1}{t}\right)X > 1 \\ &\iff 1 + \frac{2}{t-1} > \left(-\frac{t+1}{t}\right)X > 1. \end{aligned}$$

Letting $Y = \left(-\frac{t+1}{t}\right)X$ we conclude that

$$\text{Prob}\{|X| > t|X + 1|\} = \text{Prob}\left\{1 < Y < 1 + \frac{2}{t-1}\right\}. \quad (7)$$

Since $Y = \left(-\frac{t+1}{t}\right) X$, $Y \sim N(\mu_Y, \varsigma_Y^2)$ where

$$\mu_Y = \left(-\frac{t+1}{t}\right) \mu \quad \text{and} \quad \varsigma_Y^2 = \left(-\frac{t+1}{t}\right)^2 \varsigma^2. \quad (8)$$

We now apply Lemma 2 to Y with $\varepsilon = \frac{2}{t-1}$ to obtain

$$\text{Prob}\left\{1 < Y < 1 + \frac{2}{t-1}\right\} \leq \left(\frac{\mu_Y + \varsigma_Y}{\varsigma_Y}\right) \left(\frac{1}{t-1}\right) \sqrt{\frac{2}{\pi}}. \quad (9)$$

Combining (7), (8) and (9) the proof is done. \square

The following proposition is a variation of a classical result for positive random variables (cf. [2, Proposition 2]).

Proposition 2. *Let $k, H > 0$ and $X > 1$ be a random variable satisfying $\text{Prob}\{X > t\} \leq \frac{k}{t-H}$ for all $t > k + H$. Then, for all $\beta > 1$,*

$$\mathbb{E}(\log_\beta(X)) < \log_\beta(k + H) + \frac{1}{\ln \beta}.$$

Proof. We have

$$\begin{aligned} \mathbb{E}(\log_\beta(X)) &= \int_0^\infty \text{Prob}\{\log_\beta(X) > s\} ds = \int_0^\infty \text{Prob}\{X > \beta^s\} ds \\ &= \int_0^{\log_\beta(k+H)} \text{Prob}\{X > \beta^s\} ds + \int_{\log_\beta(k+H)}^\infty \text{Prob}\{X > \beta^s\} ds \\ &\leq \log_\beta(k + H) + \int_{\log_\beta(k+H)}^\infty \text{Prob}\{X > \beta^s\} ds. \end{aligned}$$

Since $\text{Prob}\{X > t\} \leq \frac{k}{t-H}$ it follows that

$$\mathbb{E}(\log_\beta(X)) - \log_\beta(k + H) \leq k \int_{\log_\beta(k+H)}^\infty \frac{dt}{\beta^t - H}. \quad (10)$$

Let $u = \frac{H}{\beta^t - H}$ so that $du = -\ln \beta(u + u^2)dt$. Then, changing variables in (10), we obtain

$$\mathbb{E}(\log_\beta(X)) - \log_\beta(k + H) \leq -\frac{k}{H \ln \beta} \int_{\frac{H}{k}}^0 \frac{du}{1 + u} = \frac{k}{H \ln \beta} \ln \left(1 + \frac{H}{k}\right).$$

The proof is complete since $\ln(1 + x) < x$ for all $x > -1$. \square

4 Computation of determinants

In this section we consider the problem of computing the determinant. Taking $F(A) = \det(A)$ in (1) we obtain the componentwise condition number $\mathbf{c}^{\det}(A)$ for this problem. Our main result for this quantity is the following.

Theorem 2. *Let $S \subset [n]^2$ be admissible, $\bar{A} \in \mathcal{M}_S$ with $\|\bar{A}\|_{\max} \leq 1$, $\sigma > 0$ and $A \sim N_S(\bar{A}, \sigma^2 \text{Id})$. Then, for any real number $t > |S|$,*

$$\text{Prob}\{\mathbf{c}^{\det}(A) > t\} < \left(\frac{1+\sigma}{\sigma}\right) \left(\frac{|S|^2}{t-|S|}\right) \sqrt{\frac{2}{\pi}}$$

and, for all $\beta > 1$,

$$\mathbb{E}(\log_{\beta}(\mathbf{c}_{\det}(A))) < \log_{\beta} \left(\frac{1+\sigma}{\sigma}\right) + 2 \log_{\beta} |S| + \frac{1.03}{\ln \beta}.$$

For the proof of this theorem we will make use of the following characterization of $\mathbf{c}^{\det}(A)$ (see [2, Lemma 1.1] for a proof). Denote by γ_{ij} the entry of A^{-1} on the i th row and j th column. Then, for any matrix $A \in \mathcal{M} \setminus \Sigma$,

$$\mathbf{c}^{\det}(A) = \sum_{i,j \in [n]} |a_{ij} \gamma_{ji}|. \quad (11)$$

Proof of Theorem 2. Without loss of generality, we may assume that $(1, 1) \in S$ so that $a_{11} \sim N(\bar{a}_{11}, \sigma^2)$. For a time to come we consider all entries of A except a_{11} to be fixed. Let A_{ij} be the matrix obtained by removing from A the i th row and j th column. By Cramer's rule, $\gamma_{11} = \frac{\det(A_{11})}{\det(A)}$ and therefore, for $t > 1$,

$$\text{Prob}\{|a_{11} \gamma_{11}| > t\} = \text{Prob}\{|a_{11} \det(A_{11})| > t |\det(A)|\}.$$

Expanding $\det(A)$ by the first column of A this equality becomes

$$\text{Prob}\{|a_{11} \gamma_{11}| > t\} = \text{Prob}\left\{|a_{11} \det(A_{11})| > t \left| \sum_{i=1}^n (-1)^{i+1} a_{i1} \det(A_{i1}) \right|\right\}$$

and letting

$$X := \frac{a_{11} \det(A_{11})}{\sum_{i=2}^n (-1)^{i+1} a_{i1} \det(A_{i1})}.$$

this equality becomes

$$\text{Prob}\{|a_{11} \gamma_{11}| > t\} = \text{Prob}\{|X| > t |X + 1|\}. \quad (12)$$

Since all entries of A , except a_{11} are fixed (and $a_{11} \sim N(\bar{a}_{11}, \sigma^2)$), we have $X \sim N(\mu, \varsigma^2)$, where

$$\mu = \frac{\bar{a}_{11} \det(A_{11})}{\sum_{i=2}^n (-1)^{i+1} a_{i1} \det(A_{i1})} \quad \text{and} \quad \varsigma = \left| \frac{\sigma \det(A_{11})}{\sum_{i=2}^n (-1)^{i+1} a_{i1} \det(A_{i1})} \right|.$$

In particular,

$$\frac{|\mu| + \varsigma}{\varsigma} = \frac{|\bar{a}_{11}| + \sigma}{\sigma} \leq \frac{1 + \sigma}{\sigma} \quad (13)$$

the last since $\|\bar{A}\|_{\max} \leq 1$. By Proposition 1, and Equations (12) and (13), we have

$$\text{Prob}\{|a_{11}\gamma_{11}| > t\} \leq \left(\frac{1 + \sigma}{\sigma} \right) \left(\frac{1}{t - 1} \right) \sqrt{\frac{2}{\pi}}.$$

This inequality holds for all fixed values of $a_{12}, a_{13}, \dots, a_{nn}$. Therefore, it holds as well when all entries of A are random (as described in Section 2.3). We can show in the same manner that, for all $(i, j) \in S$,

$$\text{Prob}\{|a_{ij}\gamma_{ji}| > t\} \leq \left(\frac{1 + \sigma}{\sigma} \right) \left(\frac{1}{t - 1} \right) \sqrt{\frac{2}{\pi}}. \quad (14)$$

We now recall that, for all $(i, j) \notin S$, $a_{ij} = 0$. Hence, by using (11), for $t > |S|$,

$$\begin{aligned} \text{Prob}\{c^{\det}(A) > t\} &= \text{Prob}\left\{ \sum_{(i,j) \in [n]^2} |a_{ij}\gamma_{ji}| > t \right\} \\ &= \text{Prob}\left\{ \sum_{(i,j) \in S} |a_{ij}\gamma_{ji}| > t \right\} \\ &\leq \sum_{(i,j) \in S} \text{Prob}\left\{ |a_{ij}\gamma_{ji}| > \frac{t}{|S|} \right\} \\ &\leq \sum_{(i,j) \in S} \left(\frac{1 + \sigma}{\sigma} \right) \left(\frac{|S|}{t - |S|} \right) \sqrt{\frac{2}{\pi}} \quad [\text{by (14)}] \\ &= \left(\frac{1 + \sigma}{\sigma} \right) \left(\frac{|S|^2}{t - |S|} \right) \sqrt{\frac{2}{\pi}}. \end{aligned} \quad (15)$$

Combining Equation (15) and Proposition 2 we obtain

$$\begin{aligned}
\mathbb{E}(\log_\beta \mathbf{c}^{\det}(A)) &\leq \log_\beta \left(|S| + \left(\frac{1+\sigma}{\sigma} \right) |S|^2 \sqrt{\frac{2}{\pi}} \right) + \frac{1}{\ln \beta} \\
&= \log_\beta \left(\left(\frac{1+\sigma}{\sigma} \right) |S|^2 \sqrt{\frac{2}{\pi}} \left(1 + \left(\frac{\sigma}{1+\sigma} \right) \frac{1}{|S|} \sqrt{\frac{\pi}{2}} \right) \right) + \frac{1}{\ln \beta} \\
&= \log_\beta \left(\left(\frac{1+\sigma}{\sigma} \right) |S|^2 \sqrt{\frac{2}{\pi}} \right) + \log_\beta \left(1 + \left(\frac{\sigma}{1+\sigma} \right) \frac{1}{|S|} \sqrt{\frac{\pi}{2}} \right) + \frac{1}{\ln \beta} \\
&\leq \log_\beta \left(\left(\frac{1+\sigma}{\sigma} \right) |S|^2 \sqrt{\frac{2}{\pi}} \right) + \frac{1}{\ln \beta} \left(\frac{\sigma}{1+\sigma} \right) \frac{1}{|S|} \sqrt{\frac{\pi}{2}} + \frac{1}{\ln \beta}.
\end{aligned}$$

The last line above is true because $\log_\beta(1+x) \leq \frac{x}{\ln \beta}$ for all $x \geq 0$. Since both σ and $|S| > 0$,

$$\begin{aligned}
\mathbb{E}(\log_\beta \mathbf{c}^{\det}(A)) &\leq \log_\beta \left(\left(\frac{1+\sigma}{\sigma} \right) |S|^2 \sqrt{\frac{2}{\pi}} \right) + \frac{1}{\ln \beta} \left(\sqrt{\frac{\pi}{2}} + 1 \right) \\
&\leq \log_\beta \left(\frac{1+\sigma}{\sigma} \right) + 2 \log_\beta |S| + \frac{1.03}{\ln \beta}. \quad \square
\end{aligned}$$

5 Matrix inversion

We next consider the problem of matrix inversion. For $k, l \in [n]$ we consider the function $F_{kl} : \mathcal{M}_S \setminus \Sigma \rightarrow \mathcal{M}$ given by $F_{kl}(A) = (A^{-1})_{kl}$. Definition (1) applied to this function yields a componentwise condition number $\mathbf{c}_{kl}^\dagger(A)$ and, recall (2), taking the maximum over $(k, l) \in [n]^2$ we obtain $\mathbf{c}^\dagger(A)$. Our main result for this quantity is the following.

Theorem 3. *Let $S \subset [n]^2$ be admissible, $\bar{A} \in \mathcal{M}_S$ such that $\|\bar{A}\|_{\max} \leq 1$, $\sigma > 0$ and $A \sim N_S(\bar{A}, \sigma^2 \mathbf{Id})$. Then, for any real number $t > 2|S|$,*

$$\text{Prob}\{\mathbf{c}^\dagger(A) > t\} = \left(\frac{1+\sigma}{\sigma} \right) \left(\frac{4n^2|S|^2}{t-2|S|} \right) \sqrt{\frac{2}{\pi}}.$$

and, for all $\beta > 1$,

$$\mathbb{E}(\log_\beta(\mathbf{c}^\dagger(A))) = \log_\beta \left(\frac{1+\sigma}{\sigma} \right) + 2 \log_\beta(n|S|) + \frac{2.65}{\ln \beta}.$$

Lemma 3. ([2, Lemma 5]) For $A \in \mathcal{M} \setminus \Sigma$ and $k, l \in [n]$,

$$\mathbf{c}_{kl}^\dagger(A) \leq \mathbf{c}^{\det}(A) + \mathbf{c}^{\det}(A_{lk}). \quad \square$$

Proof of Theorem 3. Almost certainly, $A \in \mathcal{M} \setminus \Sigma$. Hence, by Lemma 3, we have, for all $k, l \in [n]$,

$$\begin{aligned} \text{Prob}\{\mathbf{c}_{kl}^\dagger(A) > t\} &\leq \text{Prob}\{\mathbf{c}^{\det}(A) + \mathbf{c}^{\det}(A_{lk}) > t\} \\ &\leq \text{Prob}\left\{\mathbf{c}^{\det}(A) > \frac{t}{2} \text{ or } \mathbf{c}^{\det}(A_{lk}) > \frac{t}{2}\right\} \\ &\leq \text{Prob}\left\{\mathbf{c}^{\det}(A) > \frac{t}{2}\right\} + \text{Prob}\left\{\mathbf{c}^{\det}(A_{lk}) > \frac{t}{2}\right\}. \end{aligned}$$

Using Theorem 2 twice, we obtain

$$\begin{aligned} \text{Prob}\{\mathbf{c}_{kl}^\dagger(A) > t\} &\leq \left(\frac{1+\sigma}{\sigma}\right) \left(\frac{|S|^2}{\frac{t}{2} - |S|}\right) \sqrt{\frac{2}{\pi}} + \left(\frac{1+\sigma}{\sigma}\right) \left(\frac{|S|^2}{\frac{t}{2} - |S|}\right) \sqrt{\frac{2}{\pi}} \\ &= \left(\frac{1+\sigma}{\sigma}\right) \left(\frac{4|S|^2}{t - 2|S|}\right) \sqrt{\frac{2}{\pi}}. \end{aligned}$$

This inequality and the definition of $\mathbf{c}^\dagger(A)$ yield

$$\begin{aligned} \text{Prob}\{\mathbf{c}^\dagger(A) > t\} &= \text{Prob}\left\{\max_{k,l \in [n]} \mathbf{c}_{kl}^\dagger(A) > t\right\} \\ &\leq \sum_{k,l \in [n]} \text{Prob}\left\{\mathbf{c}_{kl}^\dagger(A) > t\right\} \\ &\leq \sum_{k,l \in [n]} \left(\frac{1+\sigma}{\sigma}\right) \left(\frac{4|S|^2}{t - 2|S|}\right) \sqrt{\frac{2}{\pi}} \\ &= \left(\frac{1+\sigma}{\sigma}\right) \left(\frac{4n^2|S|^2}{t - 2|S|}\right) \sqrt{\frac{2}{\pi}}. \end{aligned}$$

Finally, by Proposition 2

$$\begin{aligned}
& \mathbb{E}(\log_\beta(\mathbf{c}^\dagger(A))) \\
& \leq \log_\beta \left(2|S| + \left(\frac{1+\sigma}{\sigma} \right) (4n^2|S|^2) \sqrt{\frac{2}{\pi}} \right) + \frac{1}{\ln \beta} \\
& \leq \log_\beta \left(\left(\frac{1+\sigma}{\sigma} \right) (4n^2|S|^2) \sqrt{\frac{2}{\pi}} \left(1 + \sqrt{\frac{\pi}{8}} \right) \right) + \frac{1}{\ln \beta} \\
& = \log_\beta \left(\left(\frac{1+\sigma}{\sigma} \right) (n^2|S|^2) \right) + \log_\beta \left(\sqrt{\frac{32}{\pi}} \left(1 + \sqrt{\frac{\pi}{8}} \right) \right) + \frac{1}{\ln \beta} \\
& \leq \log_\beta \left(\left(\frac{1+\sigma}{\sigma} \right) (n^2|S|^2) \right) + \frac{2.65}{\ln \beta},
\end{aligned}$$

the second inequality due to the fact that $n, |S| \geq 1$ and $\sigma > 0$. \square

6 Linear equations solving

We finally consider linear equation solving. For $A \in \mathcal{M} \setminus \Sigma$ and $b \in \mathbb{R}^n$ we compute $x = A^{-1}b$. Thus, for $k \in [n]$, the mapping $(A, b) \mapsto x_k$ yields (always using (1)) $\mathbf{c}_k(A, b)$ and taking the maximum over $k \in [n]$ we obtain the componentwise condition number $\mathbf{c}(A, b)$ of the pair (A, b) . The following theorem is the main result in this section.

Theorem 4. *Let $S \subset [n]^2$ be admissible, $\bar{A} \in \mathcal{M}_S$ and $\bar{b} \in \mathbb{R}^n$ such that $\|\bar{A}\|_{\max} \leq 1$ and $\|\bar{b}\|_\infty \leq 1$, $\sigma > 0$, $A \sim N(\bar{A}, \sigma^2 \text{Id})$ and $b \sim N(\bar{b}, \sigma^2 \text{Id})$. Then, for any real number $t > 2|S|$,*

$$\text{Prob}\{\mathbf{c}(A, b) > t\} = \left(\frac{1+\sigma}{\sigma} \right) \left(\frac{4n|S|^2}{t-2|S|} \right) \sqrt{\frac{2}{\pi}}.$$

and, for all $\beta > 1$,

$$\mathbb{E}(\log_\beta(\mathbf{c}^\dagger(A))) = \log_\beta \left(\frac{1+\sigma}{\sigma} \right) + 2 \ln |S| + \log_\beta n + \frac{2.65}{\ln \beta}.$$

In what follows let R_k be the matrix obtained by replacing the k th column of A by b .

Lemma 4. ([2, Lemma 6]) *For any non-singular matrix A and $k \in [n]$,*

$$\mathbf{c}_k(A, b) \leq \mathbf{c}^{\det}(A) + \mathbf{c}^{\det}(R_k). \quad \square$$

Proof of Theorem 4. By Lemma 4, we have, for all $k \in [n]$,

$$\begin{aligned} \text{Prob}\{\mathbf{c}_k(A, b) > t\} &\leq \text{Prob}\{\mathbf{c}^{\det}(A) + \mathbf{c}^{\det}(R_k) > t\} \\ &\leq \text{Prob}\left\{\mathbf{c}^{\det}(A) > \frac{t}{2} \text{ or } \mathbf{c}^{\det}(R_k) > \frac{t}{2}\right\} \\ &\leq \text{Prob}\left\{\mathbf{c}^{\det}(A) > \frac{t}{2}\right\} + \text{Prob}\left\{\mathbf{c}^{\det}(R_k) > \frac{t}{2}\right\}. \end{aligned}$$

It follows from our hypothesis that $\|R_k\|_{\max} \leq 1$. We can therefore apply Theorem 2 twice to obtain

$$\text{Prob}\{\mathbf{c}_k(A, b) > t\} \leq \left(\frac{1+\sigma}{\sigma}\right) \left(\frac{4|S|^2}{t-2|S|}\right) \sqrt{\frac{2}{\pi}}$$

and, proceeding as in the proof of Theorem 3,

$$\begin{aligned} \text{Prob}\{\mathbf{c}(A, b) > t\} &= \text{Prob}\left\{\max_{k \in [n]} \mathbf{c}_k(A, b) > t\right\} \\ &\leq \sum_{k \in [n]} \text{Prob}\{\mathbf{c}_k(A, b) > t\} \\ &\leq \sum_{k \in [n]} \left(\frac{1+\sigma}{\sigma}\right) \left(\frac{4|S|^2}{t-2|S|}\right) \sqrt{\frac{2}{\pi}} \\ &= \left(\frac{1+\sigma}{\sigma}\right) \left(\frac{4n|S|^2}{t-2|S|}\right) \sqrt{\frac{2}{\pi}}. \end{aligned}$$

A last call to Proposition 2 yields the desired bound for $\mathbb{E}(\log_{\beta}(\mathbf{c}(A, b)))$. \square

7 On the accuracy of forward substitution

We arrive, at last, to the motivating theme of this paper. Theorem 1 is an immediate consequence of Theorem 4 since lower triangular matrices are sparse matrices with $S = \{(i, j) \in [n]^2 \mid i \geq j\}$. One then only needs to use that $|S| = \frac{n(n+1)}{2}$.

For the proof of Corollary 1 we use a common approach, pioneered by Wilkinson, which splits the relative error bound in the computed solution $\text{RelError}(F(a))$ as the product of two factors, one depending on the algorithm but not on the data (a backward error bound) and another depending on the data but not on the algorithm used (the condition of the data). A backward

error bound for forward substitution is shown in the following result, going back to Wilkinson [11, Ch.3,§19], which we quote, omitting some smaller details, as given in [1, Proposition 3.5].

Proposition 3. *Let $L = (l_{ij}) \in \mathbb{R}^{n \times n}$ be a nonsingular triangular matrix, $b \in \mathbb{R}^n$, and assume $\varepsilon_{\text{mach}}$ is sufficiently small (of the order of $(\log n)^{-1}$). Then, the solution \hat{x} of the system $Lx = b$ computed with forward substitution satisfies*

$$(L + E)\hat{x} = b,$$

where

$$\frac{|e_{ij}|}{|l_{ij}|} \leq (2 \log_2 n) \varepsilon_{\text{mach}}. \quad \square$$

Proposition 3 yields a backward error bound of the form $B\varepsilon_{\text{mach}}$ where $B = 2 \log_2 n$ is an expression in the dimension n of the input, independent of $\varepsilon_{\text{mach}}$.

The way such a backward error bound combines with condition to produce a bound for the loss of precision, in digits, is (see Theorem O.3 in [1])

$$\text{LoP}(F(a)) \leq \log_{10} B + \log_{10} \text{cond}^F(a) + o(1).$$

Here the $o(1)$ term is an expression tending to zero as $\varepsilon_{\text{mach}}$ does so, $\text{cond}^F(a)$ is the condition number of a and —crucially in our context— if the bound $B\varepsilon_{\text{mach}}$ is componentwise, as in Proposition 3, this condition number can be taken componentwise as well. Doing so for forward substitution and $x = L^{-1}b$ we obtain

$$\text{LoP}(x) \leq \log_{10}(2 \log_2 n) + \log_{10} c(L, b) + o(1).$$

Taking expectations on both sides and using Theorem 1 proves Corollary 1.

References

- [1] P. Bürgisser and F. Cucker. *Condition*. Forthcoming in *Grundlehren der mathematischen Wissenschaften*, Springer-Verlag.
- [2] D. Cheung and F. Cucker. Componentwise condition numbers of random sparse matrices. *SIAM J. Matrix Anal. Appl.*, 31:721–731, 2009.
- [3] N. Higham. The accuracy of solutions to triangular systems. *SIAM J. Numer. Anal.*, 26:1252–1265, 1989.

- [4] D.A. Spielman and S.-H. Teng. Smoothed analysis of algorithms. In *Proceedings of the International Congress of Mathematicians*, volume I, pages 597–606, 2002.
- [5] D.A. Spielman and S.-H. Teng. Smoothed analysis: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004.
- [6] D.A. Spielman and S.-H. Teng. Smoothed analysis of algorithms and heuristics. In *Foundations of Computational Mathematics, Santander 2005*, volume 331 of *Lecture Notes of the London Mathematical Society*, pages 274–342, 2006.
- [7] D.A. Spielman and S.-H. Teng. Smoothed analysis: An attempt to explain the behavior of algorithms in practice. *Communications of the ACM*, 52(10):77–84, 2009.
- [8] A.M. Turing. Rounding-off errors in matrix processes. *Quart. J. Mech. Appl. Math.*, 1:287–308, 1948.
- [9] D. Viswanatah and L.N. Trefethen. Condition numbers of random triangular matrices. *SIAM J. Matrix Anal. Appl.*, 19:564–581, 1998.
- [10] J. von Neumann and H.H. Goldstine. Numerical inverting matrices of high order. *Bulletin of the Amer. Math. Soc.*, 53:1021–1099, 1947.
- [11] J. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice Hall, 1963.